

ABSTRACT

A rewrite component automatically generates rewrite rules that describe how uniform resource locators (URLs) can be rewritten to reduce or eliminate different URLs that redundantly refer to the same or substantially the same content. The rewrite rules can be applied to URLs received when crawling a network to increase the efficiency of the crawl and the corresponding document index generated from the crawl.